

# **Hyperlink Analysis: Techniques and Applications**

Prasanna Desikan, Jaideep Srivastava, Vipin Kumar, and Pang-Ning Tan

Department of Computer Science,  
University of Minnesota, Minneapolis, MN, USA  
{desikan, srivastava, kumar, ptan}@cs.umn.edu

<b>ABSTRACT .....</b>	<b>0</b>
<b>1. INTRODUCTION.....</b>	<b>1</b>
1.1 HYPERLINK .....	1
1.2 HYPERLINK ANALYSIS .....	2
1.3 WEB STRUCTURE TERMINOLOGY .....	4
1.4 RELATED WORK.....	5
1.5 PAPER ORGANIZATION .....	6
<b>2. KNOWLEDGE MODELS .....</b>	<b>6</b>
2.1 GRAPH STRUCTURE MODELS .....	7
2.1.1 SINGLE NODE MODELS.....	7
2.1.2 MULTIPLE NODES MODELS .....	8
2.1.3 WHOLE GRAPH STRUCTURE.....	11
2.2 MARKOV MODELS .....	12
2.3 MAXIMAL FLOW MODELS.....	12
2.4 PROBABILISTIC RELATIONAL MODEL .....	13
2.5 OTHER MODELS .....	13
<b>3. METRICS.....</b>	<b>13</b>
3.1 METRICS FOR A SINGLE PAGE .....	14
3.1.1 HUB AND AUTHORITY SCORES .....	14
3.1.2 PAGERANK.....	15
3.1.3 STOCHASTIC APPROACH FOR LINK STRUCTURE ANALYSIS (SALSA) .....	16
3.1.4 WEB PAGE REPUTATIONS.....	18
3.2 METRICS FOR MULTIPLE PAGES.....	19
3.2.1 AVERAGE CLICKS: A MEASURE OF DISTANCE .....	19
3.2.2 INFORMATION SCENT .....	20
3.2.3 BIBLIOMETRIC METRICS.....	21
3.3 THE WHOLE WEB GRAPH .....	21
3.4 OTHER RELATED MEASURES .....	21
<b>4. ALGORITHMS .....</b>	<b>22</b>
4.1 ALGORITHMS FOR A SINGLE PAGE .....	22
4.1.1 HITS (HYPERTEXT INDUCED TOPIC SEARCH) ALGORITHM.....	22
4.1.2 PAGERANK ALGORITHM .....	24
4.2 ALGORITHMS FOR MULTIPLE PAGES.....	24
4.2.1 MAXIMAL FLOW ALGORITHM .....	25
<b>5. ANALYSIS SCOPE.....</b>	<b>26</b>
<b>6. APPLICATIONS OF HYPERLINK ANALYSIS.....</b>	<b>28</b>
6.1 TOPIC DISTILLATION .....	28
6.2 WEB PAGE CATEGORIZATION.....	29
6.3 IDENTIFICATION OF WEB COMMUNITIES.....	29
6.4 WEB CRAWLING.....	30
6.5 WEB USAGE BASED APPLICATIONS .....	32
<b>7. METHODOLOGY FOR APPLYING HYPERLINK ANALYSIS .....</b>	<b>32</b>
7.1 CLASSIFICATION OF APPLICATIONS USING HYPERLINK ANALYSIS .....	33
7.2 HYPERLINK ANALYSIS METHODOLOGY .....	33
<b>8. CONCLUSIONS .....</b>	<b>35</b>
<b>9. ACKNOWLEDGEMENTS.....</b>	<b>36</b>
<b>REFERENCES .....</b>	<b>36</b>

## Abstract

The concept of hyperlinks was introduced with the invention of hypertext. Though originally conceived as a mechanism to dynamically link a citation to its actual source, the recent past has seen its usage grow in ways that could not have been conceived just a few years ago. Hyperlink Analysis is the name given to a collection of techniques that have emerged to analyze the hyperlink structure that exists in the Web. The analysis can be for a wide variety of purposes, ranging from ranking pages returned from a web search engine to understanding the social dynamics behind the usage of the Web as a whole. Although this field is relatively new, rapid interest has led to the development of a significant body of literature, reporting on emerging techniques for hyperlink analysis as well as experience in their usage. As is to be expected of any new area, while a number of creative ideas have emerged, the interconnections between them are not clearly evident. Often solutions to the same core problems have been arrived in widely different ways – and reported as such – based on the respective perspectives of the investigators. We believe the reason for this is the lack of a systematic cataloging of the existing literature, which makes the similarities and complementarities of various approaches clearer. The goal of our effort is to fill this gap. In this survey we introduce a taxonomy for classifying the research on hyperlink analysis. Four key dimensions, namely *knowledge models*, *metrics*, *algorithms* and *analysis scope* are identified. We describe each of these dimensions in detail, and show how they form the core components of any application of hyperlink analysis. We classify the existing literature in terms of this taxonomy, and thereby illustrate where they are similar and where they complement each other. A rather pleasing consequence of the taxonomy is that it leads naturally to a methodology for applying hyperlink analysis for an application that has been described. We conclude the survey by briefly summarizing our work and its purpose.

# 1. Introduction

Information retrieval on the World Wide Web has been one of the challenging tasks in recent years. Most early work on Information Retrieval concentrated on the content portion of the hypertext, and little attention was paid to the hyperlinks connecting the various documents. Google [1] was one of the earliest search engines that exploited the hyperlink information to improve the quality of search. The effectiveness of Google and its popularity has increased interests in using hyperlinks to mine information from the World Wide Web. In this paper we describe the nature of a hyperlink and how it can be used as an additional instrument in effectively mining the World Wide Web.

## 1.1 Hyperlink

A hyperlink is a structural unit that connects two Web pages as shown in Figure 1. This connection is realized by inserting a hyperlink at the desired point in the source page. The hyperlink contains the URL of the destination page.

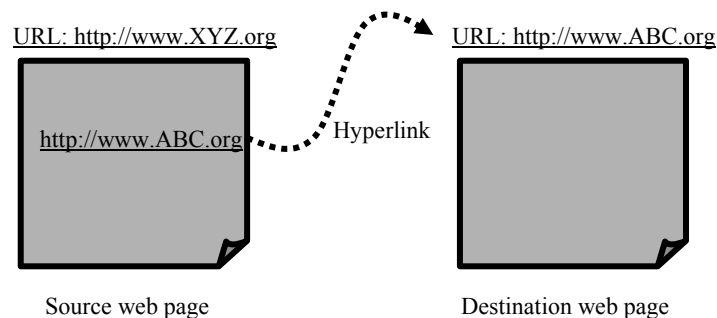


Figure 1. Hyperlink

When a user browsing the source page clicks on the hyperlink, the Web browser interprets this as a request to fetch the page referenced by the hyperlink<sup>1</sup>. Hyperlinks can be used for purely navigational purposes or to point to other pages that are related to the topic of the page containing the hyperlink. Hyperlinks are similar to the citations that form links between research papers in scientific literature. A key difference lies in the fact that they do not have a temporal dimension – in the sense that citations in a paper that has already been published cannot be altered. Also, citations in a paper cannot point

---

<sup>1</sup> This execution semantics of a hyperlink has been universally defined, and every browser must implement it.

to papers that have been published later than the paper itself. These issues have been discussed in [2]. In the past few years, hyperlinks have been used in a wide variety of ways, with many different semantics, usually based on the application. This widespread use of hyperlinks has made *hyperlink analysis* an emerging and important area of research.

## 1.2 Hyperlink Analysis

Hyperlink Analysis by itself is a part of bigger research area - Web Mining, which can be described as the process of applying data mining techniques to extract useful information from Web data. The kinds of data that can be collected and used in Web Mining analysis include *content data*, *structure data*, and *usage data* [3]. As a result, the field of Web Mining can be broadly divided into three distinct categories, according to the kinds of data to be mined [3,4]:

1. **Web Content Mining:** Web Content Mining is the process of extracting useful information from the contents of Web documents. Content data corresponds to the collection of facts a Web page was designed to convey to the users. It may consist of text, images, audio, video, or structured records such as lists and tables. Research activities in this field also involve using techniques from other disciplines such as Information Retrieval (IR) and natural language processing (NLP).
2. **Web Structure Mining:** The structure of a typical Web graph consists of Web pages as nodes, and hyperlinks as edges connecting between two related pages. In addition, the content within a Web page can also be organized in a tree-structured format, based on the various HTML and XML tags within the page. Thus, Web Structure Mining can be regarded as the process of discovering structure information from the Web. This type of mining can be performed either at the (intra-page) document level or at the (inter-page) hyperlink level.
3. **Web Usage Mining:** Web Usage Mining is the application of data mining techniques to discover interesting usage patterns from Web data, in order to understand and better serve the needs of Web-based applications [3]. Usage data captures the identity or origin of Web users along with their browsing behavior at

a Web site. Some of the typical usage data collected at a Web site include IP addresses, page references, and access time of the users.

Figure 2 below presents a high-level taxonomy of the various research activities in Web Mining. For Web Structure Mining, we can divide this field further into two sub-categories, namely document structure analysis (the structure of a document such as the Document Object Model) and link type analysis (structure due to the links referring to within a document or those referring to other documents). As the figure suggests, in Hyperlink Analysis, we concentrate only on the information that can be extracted from the inter-document link structure. However, hyperlink analysis can be enriched by information extracted from document structure analysis, Web Content Mining or Web Usage Mining. For example, Henzinger defines *Link Analysis* in [5] as the area of information retrieval using hyperlinks as the source of information. Hyperlinks provide structural information which, coupled with Web content, can be used to mine useful information from the Web and to measure the quality of information.

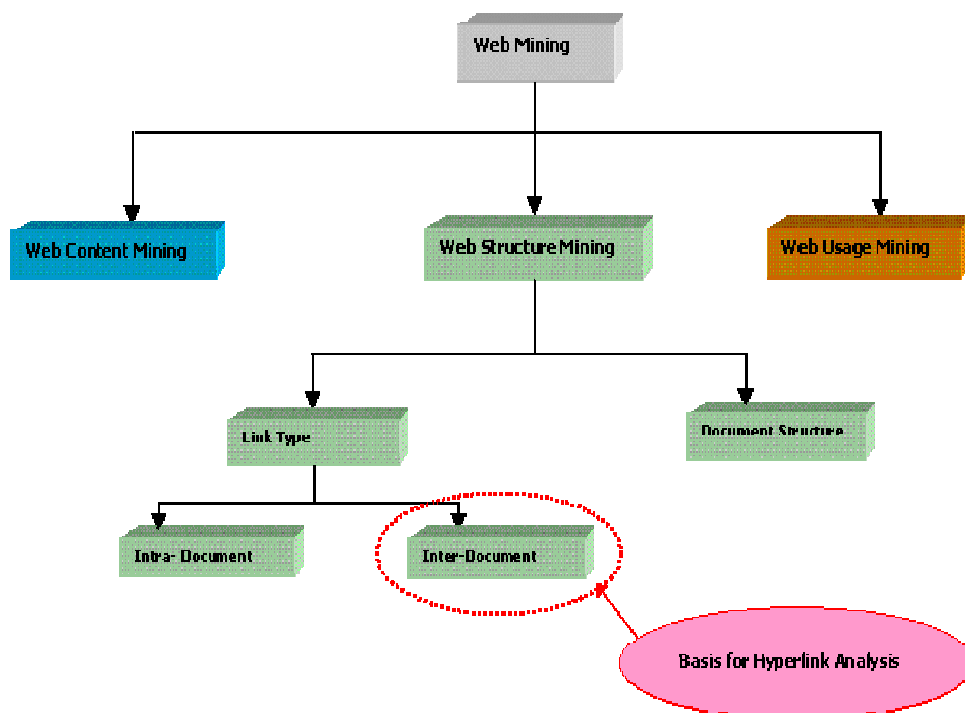


Figure 2. Taxonomy of Web Mining with Web Structure mining expanded further to explain the scope of Hyperlink Analysis.

Hyperlink analysis can be used for a variety of purposes. Some of the main uses are:

- Measuring the extent of support that the ideas and statements on a page provide for a particular topic. This information also helps to rank Web pages according to their relative importance.
- Characterizing the Web graph structure by examining the various graph patterns, such as co-citations, co-references, bi-partite graphs etc.
- Serving as an effective tool in classifying Web pages according to various topics and functionalities.
- Improving the efficiency of *crawling* by identifying the relative importance of pages that need to be crawled first.
- When combined with usage statistics, hyperlink analysis can be used for predicting user-browsing behavior and help the user to surf the Web better.

### 1.3 Web Structure Terminology

The Web as a whole can be modeled as a directed graph containing a set of nodes and directed edges between them. Broder et al [7] studied the web graph and described some of the basic terminology necessary for a web graph model. The nodes represent the Web pages and the directed edges are the hyperlinks. We now define a set of terms that are frequently used to describe the Web graph structure and other more abstract concepts about the Web.

*Web-graph*: A directed graph that represents the Web.

*Node*: Each Web page is a node of the Web-graph.

*Link*: Each hyperlink on the Web is a directed edge of the Web-graph.

*Indegree*: The indegree of a node,  $p$ , is the number of distinct links that point to  $p$ .

*Outdegree*: The outdegree of a node,  $p$ , is the number of distinct links originating at  $p$  that point to other nodes.

*Directed Path*: A sequence of links, starting from  $p$  that can be followed to reach  $q$ .<sup>2</sup>

*Shortest Path*: Of all the paths between nodes  $p$  and  $q$ , which has the shortest length, i.e. number of links on it.

*Diameter*: The maximum of all the shortest paths between a pair of nodes  $p$  and  $q$ , for all pairs of nodes  $p$  and  $q$  in the Web-graph.

---

<sup>2</sup> A link can be traversed in only one direction, i.e. from its source to its destination

*Average Connected Distance:* Average of the lengths of the shortest paths from node  $p$  to node  $q$ , for all pairs of nodes  $p$  and  $q$  [6]. Broder et al. [7] observed that this definition could result in an infinite average connected distance, if there is at least one pair of nodes  $p$  and  $q$  that have no existing path between them. And they proposed a revised definition: “the average connected distance is the expected length of the shortest path, where expectation is uniform choices from a set of all ordered pairs,  $(p,q)$  such that there exists a path from  $p$  to  $q$ ”

#### 1.4 Related Work

The past few years have seen a growing interest in the research in Web Mining. In [8], Etzioni first suggested that the Web can be seen as composed of documents with structured information, and features can be extracted from them for effective mining of knowledge from the Web. The mining process was divided into three phases, namely *resource discovery*, *information extraction* and *generalization*. Cooley et al [9], classified Web data into three categories, namely *content*, *structure* and *usage*. Srivastava et al [3] surveyed the various research activities in Web Usage Mining and identified a number of applications for it. Kosala and Blockeel in their survey [4] classified Web mining research into three categories, namely *Web Content Mining*, *Web Structure Mining*, *Web Usage Mining*. They also view Web mining from the perspective of agent based paradigms such as intelligent agents and software agents that perform data mining tasks. [3,4] address the various research issues in *Web Usage Mining* and serve as a good survey for the field. Chakrabarti [10] compared the different data mining and statistical techniques that have been applied to the hypertext documents and their applications in the Web domain. Efe et al [2] discuss the importance of links and the interesting graph patterns that are formed. They also give a overview of couple of link-based metrics. Henzinger [5] describes two successful link based ranking methods and briefly describes the possible areas of research in “link analysis”. Our survey concentrates on the analysis carried out using the information provided by the inter-document link structure with or without combining the document structure, Web content or Web usage information. We describe the basic dimensions required for hyperlink analysis and how they relate to the applications.



## 1.5 Paper Organization

Web Mining as a whole is a vast area of research, for which a high-level taxonomy is presented in Figure 1. In this survey we concentrate only on the work related to using hyperlinks for extracting useful information from the Web. More specifically, we concentrate on the inter-document link structures and a combination of it with Web content, Web usage, or the document structure of a Web page. We do not include research that uses purely content or usage data in its analysis. Any research on Hyperlink Analysis can be analyzed along the following four dimensions:

- *Knowledge Models*: The underlying representations that forms the basis to carry out the application specific task. The representation could be based on graph models, flow models or probabilistic models
- *Metrics*: The metrics used to measure Web page properties such as quality, relevance, and structural properties like distance between pages or the properties of the whole Web like its diameter.
- *Algorithms*: The procedures followed based on the underlying knowledge models to compute a metric or measure or to identify other Web graph related properties.
- *Analysis Scope*: The scope of the analysis specifies if the task is relevant to a single page or set of pages or the entire Web.

This survey is organized along the dimensions outline above, which we describe in detail in the following sections. In Section 2, we discuss the *knowledge models* used. Section 3 surveys the various *metrics* proposed for determining Web page and Web structure properties. In section 4 we survey the various *algorithms* proposed in the literature. *Analysis scope* and its role are discussed in section 5. The various applications of hyperlink analysis are discussed in section 6. In section 7 we describe a methodology for hyperlink analysis and present the different projects done in the area based on the dimensions proposed in section 1. Section 8 concludes the paper with directions for future research.

## 2. Knowledge Models

Most research in Hyperlink Analysis starts with a basic model upon which different measures are applied and the targeted application objective is achieved by a more specific

computation technique or algorithm. These models either relate to the basic information unit or the process that focus on the application. Different kinds of models, based on graph structures, statistical methods or network-flow have been proposed.

## 2.1 Graph Structure Models

In this section we discuss the various graph structures that represent certain concepts and serve as information units while mining the Web. Graph structures comprise of single node, multiple nodes or the whole set of nodes that constitute the graph. The following graph structures have been proposed for hyperlink analysis:

### 2.1.1 Single Node Models

Single Node Models are graph structures consisting of a single node and the links pointing to or away from it.

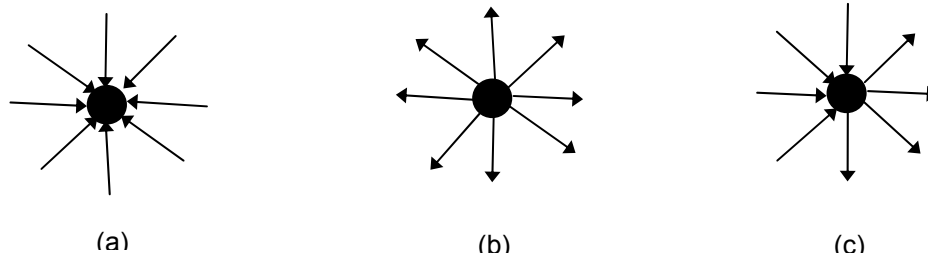


Figure 3. Single Node Models. (a) represents a pure *authority* page. (b) represents a pure *hub* page. (c) represents more a typical Web page that will have both a *hub* score and an *authority* score associated with it.

*Authority:* An *authority* page is a Web page that is pointed to by a large set of other Web pages.

*Hub:* A *hub* page is a Web page that points to a large set of other Web pages. A *good hub* is a one that points to many good *authorities*, while a good *authority* is one that is pointed to by many good *hubs*. The notion of *hubs* and *authorities* was first introduced by Kleinberg in [29]. Single page models are often used to determine the quality of a Web page [11, 12, 13].

### 2.1.2 Multiple Nodes Models

Multiple Node Models deal with graph structures that contain a set of nodes and the links that connect them. Some these graph structures or patterns have also been discussed by Efe et al. [2]. We describe below these models and the concepts they reflect:

*Direct Reference*: A direct reference refers to a concept where a node A is pointed to directly by an adjacent node B. In Figure 3(a), 'B' is *directly referred* by 'A' indicating that 'A' and 'B' may address a common topic and may be related.

*Indirect Reference*: An indirect reference refers to a concept where node A is pointed to or referred directly by an adjacent node B and node B is pointed to or referred directly by another adjacent node C, then node A is said to be indirectly referred by node C. In Figure 4(b), 'A' directly refers 'B' and 'B' directly refers 'C'. Thus 'A' *indirectly refers* 'C' indicating that 'A' and 'C' could be related.

*Mutual Reference*: When two nodes A and B point to each other directly, then they are said to mutually-reference each other. This also indicates a strong relevance between the two pages. In Figure 4(c), 'A' and 'B' are said to *mutually refer* each other.

*Co-Citation*: When a node A points to two other nodes B and C, then node A is said to be co-citing node B and node C. On the Web, such co-citation intuitively could indicate a similarity between page B and page C. In Figure 4(d), 'A' is co-citing 'B' and 'C'. Thus, it is possible that 'B' and 'C' have some similarity.

*Co-Reference*: When two nodes B and C point to a node A, then node A is said to be co-referenced by node B and node C. On the Web, such co-citation intuitively indicates a possible similarity between page B and page C. In Figure 4(e), 'C' is co-referenced by 'A' and 'B' suggesting possible relatedness between 'A' and 'B'.

*Directed Bipartite Graph*: A graph whose node set can be partitioned into two disjoint sets  $F$  and  $C$ , where every directed edge in the graph is from a node  $u$  in  $F$  to a node  $v$  in  $C$ .

*Complete Bipartite Graph*: A bipartite graph that contains all possible edges between a vertex of  $F$  and a vertex of  $C$ .

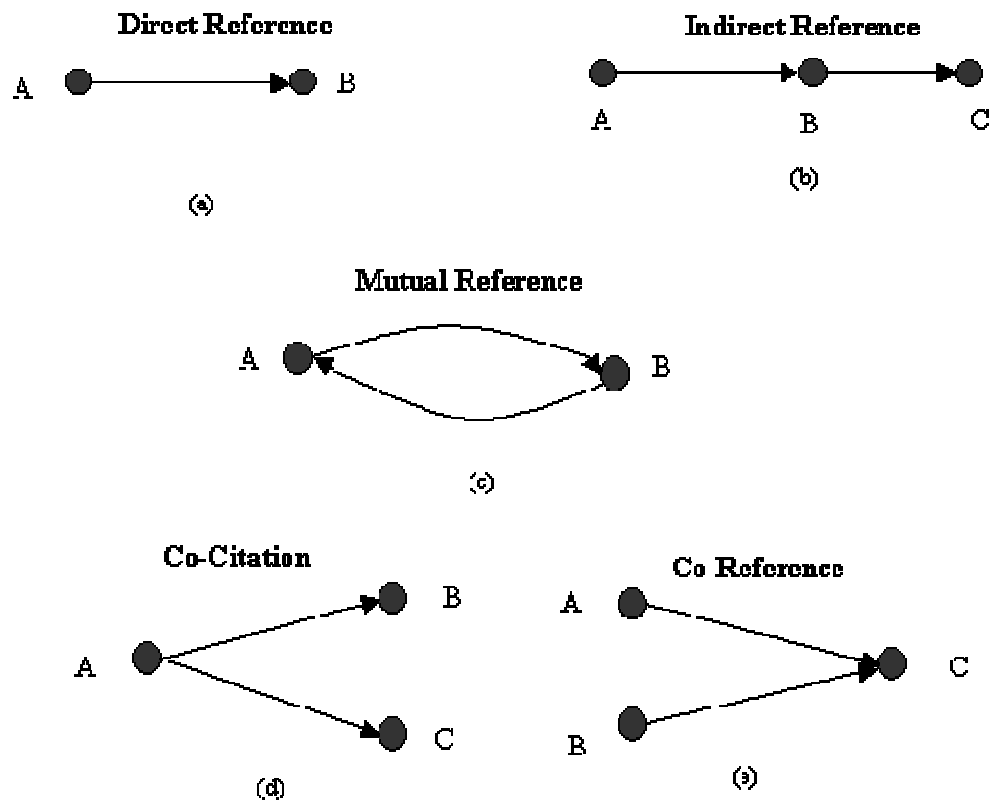
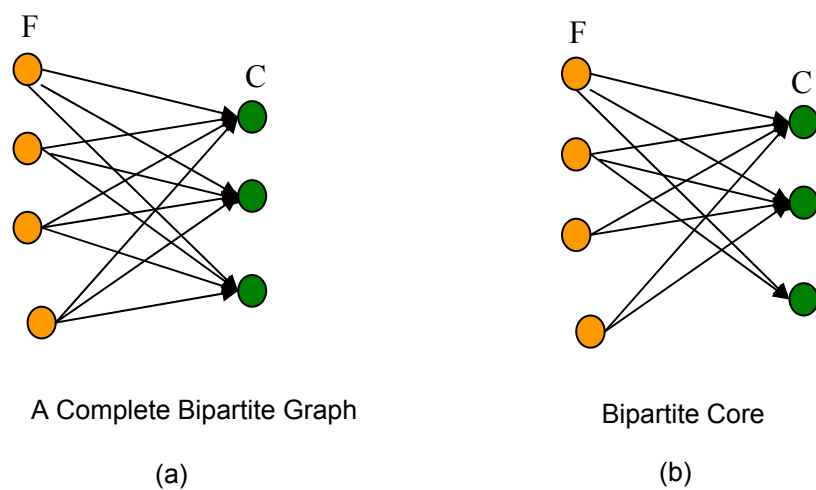


Fig 4. Multiple Node Models with simple structures. They have also been discussed in [2] as graph patterns.



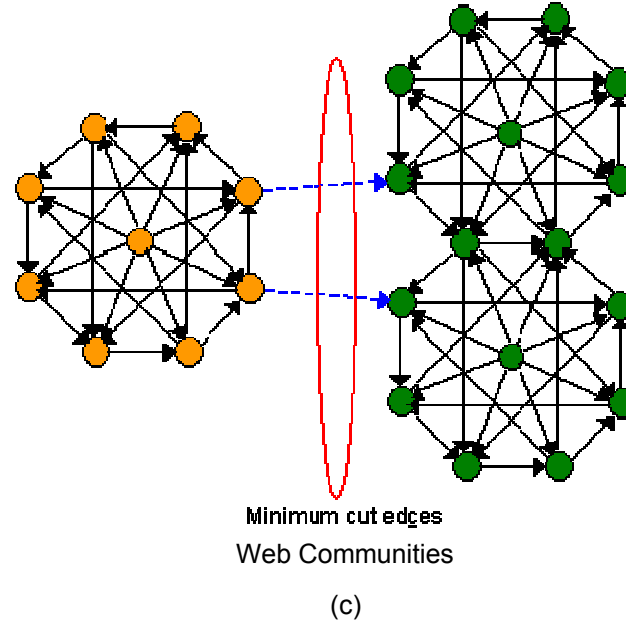


Fig 5. Multiple Node Models with more complex structures. In (a) the set of nodes on the left side are called *fans* or *hubs* and the set of nodes on the right side are called the *centers* or *authorities*. In (b) the last two nodes on the left side do not connect to all nodes on the right side. The first two nodes on the left side and the first two nodes on the right side form a complete bipartite graph. In (c) the structures on the left and right side of the minimum cut edges represent Web Communities as defined by Flake et al [15]

*Bipartite Core*: A Core  $(i, j)$  is a complete directed bipartite sub-graph with at least  $i$  nodes from  $F$  and at least  $j$  nodes from  $C$ . With reference to the Web graph, the  $i$  pages that contain the links are referred to as ‘*fans*’ and the  $j$  pages that are referenced are the ‘*centers*’. From a conceptual point of view ‘*fans*’ and ‘*centers*’ in a Bipartite Core are basically the *Hubs* and *Authorities*. For a set of pages related to a topic, bipartite core can be found that represents the *Hubs* and *Authorities* for the topic. *Hubs* and *Authorities* are important since they serve as good sources of information for the topic in question.

*Community*: Community is a core of central authoritative pages linked together by hub pages [14]. It has also been defined as a collection of Web pages such that each member

node has more hyperlinks (in either direction) within the community than outside of the community [15].

### 2.1.3 Whole Graph Structure

*Bow-Tie Model:* Broder et al. [7, 16] proposed the “bow-tie” model of the Web graph. They discuss in detail the properties of the Web graph and the different measurements, methods, and models applied to the Web graph. The “*bow-tie*” model consists of one central strongly connected component (SCC), a second weak component (IN) that has links pointing from nodes in it to the strongly connected component, and a third weak component (OUT) that has links pointing from the strongly connected component to nodes in it.

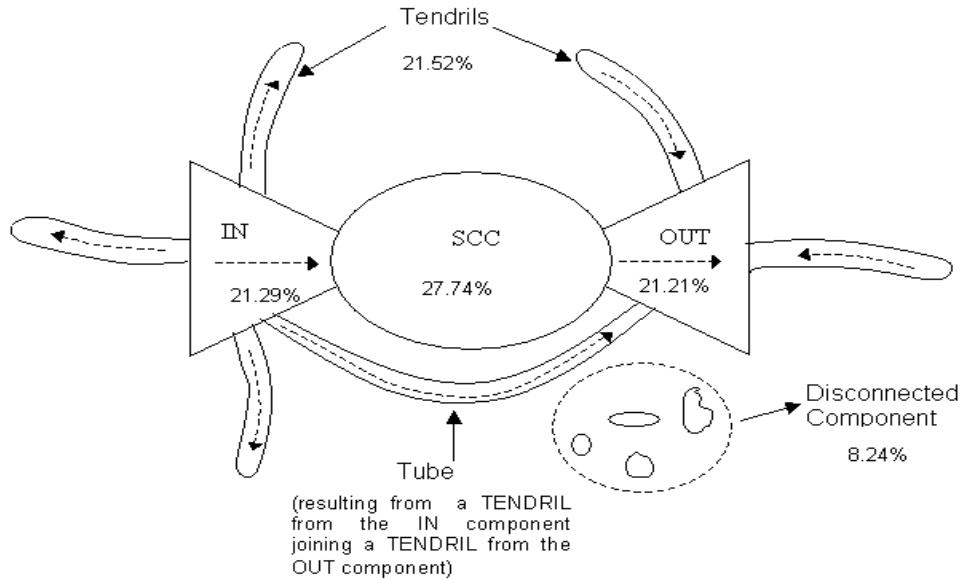


Figure 6. The “bow-tie” model proposed by Broder et al. in [7]

There are some disconnected components that are not linked to these main structures. Asets of nodes called TENDRILS start from either IN and are not connected to the SCC, or are connected to OUT independently of the other two components. TUBES can be viewed as the TENDRILS from the IN component joining the TENDRILS from the OUT component. So they are all nodes that are connected from the IN component to the OUT

component and do not belong to the SCC. Broder et al. suggest that further work could be done to develop mathematical models for evolving graphs.

## 2.2 Markov Models

The underlying principle of an ‘m’ order Markov chain is that given the current state of a system, the evolution of the system in the future depends only on the present state and the past ‘m-1’ states of the system. First order Markov models have been used to model the browsing behavior of a typical user on the Web. PageRank [11] and Randomized HITS [17] use the random walk process based on Markov model. The user randomly chooses to either jump to a new page or to follow a link – *outlink* in case of PageRank, and *inlink* or *outlink* depending on the time-step in case of Randomized HITS approach. Other approaches, e.g. SALSA [12] have also incorporated the Markovian random walk. Zhu et al [18] use Markov chains to predict links for adaptive Web sites. The modeling of a Web surfer, that essentially involves traversing a link, based on Markov models has been used significantly in hyperlink analysis.

## 2.3 Maximal Flow Models

The  $s - t$  maximal flow problem can be described thus: Given a graph  $G = (V, E)$  whose edges are assigned positive flow capacities, and with a pair of distinguished nodes  $s$  and  $t$ , the problem is to find the maximum flow that can be routed from the  $s$  to  $t$ .  $s$  is known as the *source* node and  $t$  as the *sink* node. Of course, the flow must strictly adhere to the constraints that arise due to the edge capacities. Ford and Fulkerson [19] proposed that the maximal flow is equivalent to a “minimal cut” – that is the minimum number of edges that need to be cut from the graph to separate the source  $s$  from sink  $t$ . A number of algorithms have been proposed to solve the problem, e.g. the shortest augmentation path problem [20] or the one by [21]. This approach has been used by Flake et al [16], [22] to identify “Web communities”, which are characterized by the set of pages that are linked to more Web pages within the “community than to Web pages outside the “community”, as depicted in Figure 5(c).

## 2.4 Probabilistic Relational Model

A *probabilistic relational model (PRM)* is basically a combination of the relational model with the Bayesian belief network. The relationships among attributes within a class and the relationship of attributes across different classes can be modeled by assigning different probability distributions. Getoor et al [23] treat Web documents and Links as entities and the assigning the relationships between them. A Web document entity would have attributes like: *hub*, *category*, *words* etc. And a Link entity would have an attribute like: *exists* – to model the relationship with the Web document entity. The value of the link attributes states if a link to the document exists or not. Starting with pre-assigned probability distributions for the *category* attributes of a document entity, they used the Bayesian approach and the belief propagation method to classify documents.

## 2.5 Other Models

We briefly mention some of the other models that have been used. Cohn et al [24] develop a '*probabilistic factored model*' to identify "authoritative" documents by determining the conditional probability,  $P(c/z)$  that a document 'c' is a cited given a topic category 'z'. Finally [25] describes *Agora* pages that are pages linked from multiple-pages, each of which are the highest ranked in their community according to Google's PageRank. The intent is to identify the emergence of future communities. There are other slight modifications of *hubs* and *authorities* and the *PageRank* e.g. [17, 12, 26, 27, 28] that are not discussed here.

## 3. Metrics

Hyperlinks can be used to define standards for measuring the properties of an individual Web page, a group of pages or the whole Web structure. Hyperlink analysis has been used as an effective tool to measure authority of Web pages on topics, computing Web page reputations and measuring distance between Web pages. Different methods have been proposed to identify the quality of Web pages using hyperlinks. While some metrics are based on the principle of random walks on a Web graph, others are based on Web graph structures like complete bipartite graph and bipartite core. Yet others have followed probabilistic and stochastic approaches. The metrics also differ from the fact



that they are either query dependent or query independent. In this section, we discuss in detail some of the more popular and interesting metrics that make use of the hyperlink information. Hyperlinks have been useful in measuring the properties of a single *Web Page*, a set of *Web Pages* or the *Web Graph* as a whole. *Average Clicks* is a measure developed to find the distance between Web pages using the hyperlink structure. We structure our discussion on these measures and metrics based on if they apply to a single page, multiple pages or the whole Web.

### 3.1 Metrics for a Single Page

*PageRank* [11], *HITS* [29], *SALSA* [12] and *Web Page Reputations* [13] measure the quality of individual Web pages. *PageRank* is a metric, while *HITS* and *SALSA* viewed from a bigger picture are names of approaches that use their own metric to determine the quality of a Web page. *Web Page Reputation* is a derived metric based on *PageRank* to measure the “reputation” of a page for a given topic. The following sections describe these metrics that apply to individual pages.

#### 3.1.1 Hub and Authority Scores

*Hubs* and *Authorities*, as mentioned earlier, together form a bipartite graph, with the directed edges formed by the *hubs* pointing to the *authorities*. The *hub* and authority scores computed for each Web page indicate the extent to which the Web page serves as an “authority” on a topic or as a “hub” that points to good “authority” pages. The *hub* and *authority* scores for a page are not based on a single formula, but are computed for a set of pages related to a topic using the HITS algorithm [29] described in section 4.1.1. We briefly give an overview of the procedure to obtain these scores. First a query is submitted to a search engine and a set of relevant documents is retrieved from a search engine. This set is called the ‘*root set*’. The ‘*root set*’ is then expanded by including Web pages that point to Web pages in the ‘*root set*’ and are pointed by the Web pages in the ‘*root set*’. This whole new set is called the ‘*Base Set*’. An adjacency matrix,  $A$  is formed such that if there exists at least one hyperlink from page  $i$  to page  $j$ , then  $A_{i,j} = 1$ , else  $A_{i,j} = 0$ . HITS algorithm is then used to determine the *Hubs* and *Authorities* scores.

The HITS approach has been found in general to be successful on queries on topics that are well represented in the Web in terms of linkage density. Sometimes, when a query on a narrower topic is issued, HITS tends to return results for a more general topic or vice-versa. This problem of “topic drift” occurs because HITS tends to converge to topics that have a better density of linkage on the Web graph.

In [27] Chakrabarti et al. modified the Kleinberg’s *hub* and *authority* scores by using text-based weights in the adjacency matrix while calculating the scores. Bharath and Henzinger in [28] suggested that edge weights should be modified such that if there are  $k$  edges on a document on the first host pointing to a single document on the second host, each edge is given an *authority weight* of  $1/k$ . Similarly, if a document on a host is pointing to  $l$  documents on another host, then each edge is given a weight of  $1/l$ . This would solve the problem of “mutually reinforcing relationships” between hosts. The CLEVER project at IBM [30] has enhanced the original HITS based measures, and used it for link-based applications like Web crawling, Web page categorization and Web communities.

### 3.1.2 PageRank

PageRank is a metric for ranking hypertext documents that determines the quality of these documents. Page et al. [11] developed this metric for the popular search engine, Google [1,31]. The key idea is that a page has high rank if the sum of the ranks of its backlinks (links pointing to the page) is high. So the rank of a page depends upon the ranks of the pages pointing to it. This process is done iteratively till the rank of all the pages is determined. The rank of a page  $p$  can thus be written as:

$$PR(p) = \frac{d}{n} + (1-d) \cdot \sum_{(q,p) \in G} \frac{PR(q)}{OutDegree(q)}$$

The dampening factor  $d$  is usually set between 0.1 and 0.2. Here,  $n$  the number of nodes in the graph and  $OutDegree(q)$  is the number of hyperlinks on page  $q$ .

Intuitively, the approach can be viewed as stochastic analysis of a random walk on the Web graph. The first term in the right hand side of the equation corresponds to the probability that a random Web surfer arrives at a page  $p$  out of nowhere, i.e. he could arrive at the page by typing the URL or from a bookmark or may have a particular page as his/her homepage.  $d$  would then be the probability that a random surfer chooses to type a URL versus traversing a link. And  $1/n$  corresponds to the uniform probability that a person chooses the page  $p$  from the whole set of  $n$  pages. The second term in the right hand side of the equation corresponds to factor contributed by arriving at a page by traversing a link.  $1 - d$  is the probability that a person arrives at the page  $p$  by traversing a link. The summation corresponds to the sum of the rank contributions made by all the pages that point to the page  $p$ . The rank contribution is the PageRank of the page multiplied by the probability that a particular link on the page is traversed. So for any page  $q$  pointing to page  $p$ , the probability that the link pointing to page  $p$  is traversed would be  $1/OutDegree(q)$ , assuming all links on the page is chosen with uniform probability.

The original PageRank is a query independent approach and hence is a global ranking system. Haveliwala in [32] discusses efficient methods to scale the implementation of PageRank to large subgraphs on machines with limited capacity of memory. PageRank is also found to be very stable. The stability of PageRank and other ranking metrics is discussed in [17,26,33]. According to [17,33], as long as Web pages with high PageRank scores are not modified or perturbed (i.e. either more links are added or certain links removed), the PageRank scores resulting from perturbing or modifying any Web page will not very different from the original PageRank scores. One of the main reasons is attributed to the factor contributed by arriving at a page out of nowhere (i.e. choosing to type an URL of a page chosen at random from a uniform distribution) to the total score.

### **3.1.3 Stochastic Approach for Link Structure Analysis (SALSA)**

The SALSA approach was proposed by Lempel and Moran in [12]. It combines the theory of random walks with Kleinberg's approach of using bipartite graphs representing *hubs* and *authorities*. It is also a query dependent approach. The intuition behind this approach is if a surfer walks randomly on a sub graph generated by a issuing a query on a

topic, the surfer, with high probability, should visit the pages that have higher quality information or the pages with a high “authority” on the topic. Similarly, the surfer, with high probability, should be able to visit pages that point to these authority pages during his random walk.

We briefly describe the procedure involved in computing the metric for this approach. In the first step, a base set is constructed as is done by the HITS algorithm. Next a bipartite undirected graph,  $G$ , is built from this collection. To estimate the *hub* and *authority* scores two different kinds of walks are performed. One visiting the nodes on the *hub* side and generating a chain of *hubs*, and the other visiting the nodes on the *authority* side and generating a chain of *authorities*. While performing a random walk, the surfer starts from one side of the bipartite graph and in each step traverses two edges – the first edge to lead him to the other side of the graph and the second edge to lead him back to the original side. The Markov model applied here helps in estimating a probability that the surfer will visit the nodes on one side of the bipartite core, either the *hubs* or the *authorities* side.

The different Markov Chains generated by these two distinct walks – the *authority* chain and the *hub* chain – in turn help in generating the *authority* and *hub* scores for each Web site. The transition matrices containing the transition probabilities generated by the two Markov Chains are:

1. The Hub-Matrix,  $H$ , whose element is represented as:

$$h_{i,j} = \sum_{\{k|(i_h, k_a), (j_h, k_a) \in G\}} \frac{1}{OutDegree(i_h)} \cdot \frac{1}{InDegree(k_a)}$$

2. The Authority-Matrix,  $A$ , whose element is represented as:

$$a_{i,j} = \sum_{\{k|(k_h, i_a), (k_h, j_a) \in G\}} \frac{1}{InDegree(i_a)} \cdot \frac{1}{OutDegree(k_h)}$$

The principal eigenvectors of these matrices  $A$  and  $H$  will give identify and hubs respectively. Due to the introduction of the theory of random walk, the SALSA approach doesn't always close in on sub graphs with high linkage density and hence avoids the problem of "topic drift".

### 3.1.4 Web Page Reputations

In [13] the authors come up with the concept of "reputation" of a Web page on different topics. Similar to SALSA they perform a two-level random walk and come up with a metric to determine the authority reputation and hub reputation of a page on a topic  $t$ . The underlying principle is at each step:

- with probability  $d > 0$  the surfer jumps to a random page, or
- with probability  $(1-d)$  the surfer follows a random link forward/backward from the current page alternating in directions.

With this kind of model, they define the two kinds of reputation a Web page has on a topic  $t$ :

- **Authority Reputation** of a page  $p$  on a topic  $t$  is the probability that a random surfer looking for a topic  $t$  makes a forward visit to the page  $p$ .
- **Hub Reputation** of a page  $p$  on a topic  $t$  is the probability that a random surfer looking for a topic  $t$  makes a backward visit to the page  $p$

At each step, the random surfer picks a direction – forward or backward – with a probability equal to  $1/2$ , with a probability 'd' chooses to jumps and a probability  $1/N_t$  chooses the particular page  $p$ , where  $N_t$  is the number of pages on topic  $t$ . Hence, with a probability  $d/2N_t$  the surfer makes a forward (or backward ) visit to a page  $p$  in a random jump. The second term will be the probability that a surfer will visit a page following a link. The metrics for computing the authority and hub reputations for a page on a topic  $t$  is given by:

$A^n(p,t) \leftarrow$  probability of a forward visit to a page  $p$  when searching for a term  $t$  at step  $n$   
or in other words "*Authority Rank*" of page  $p$  on a term  $t$

$$\begin{aligned}
A^n(p, t) &= \frac{d}{2 \cdot N_t} + (1-d) \cdot \sum_{q \rightarrow p} \frac{H^{n-1}(q, t)}{\text{Out}(q)} \quad \text{if term } t \text{ appears in a page } p \\
&= (1-d) \cdot \sum_{q \rightarrow p} \frac{H^{n-1}(q, t)}{\text{Out}(q)} \quad \text{otherwise}
\end{aligned}$$

$H^n(p, t) \leftarrow$  probability of a backward visit to a page  $p$  when searching for a term  $t$  at step  $n$  or in other words “*Hub Rank*” of page  $p$  on a term  $t$

$$\begin{aligned}
H^n(p, t) &= \frac{d}{2 \cdot N_t} + (1-d) \cdot \sum_{p \rightarrow q} \frac{A^{n-1}(q, t)}{\text{In}(q)} \quad \text{if term } t \text{ appears in a page } p \\
&= (1-d) \cdot \sum_{p \rightarrow q} \frac{A^{n-1}(q, t)}{\text{In}(q)} \quad \text{otherwise}
\end{aligned}$$

### 3.2 Metrics for Multiple Pages

In this section we discuss the approaches for quantifying the properties involving two or more pages and the relationships between them. These relationships include distance between pages, co-citations, co –references etc.

#### 3.2.1 Average Clicks: A Measure of Distance

Matsuo et al. in [34] proposed that the distance measured by the number of clicks doesn’t reflect well the users’ intuition of distance. It is more likely that a user will follow a link from a page that has few links, than follow a link from a page that has many links. Hence they proposed a new measure of distance called – *average clicks*. This distance is based on the probability of clicking on a link through random surfing. With this model the length of a link in a page  $p$  is defined as:

Length of Link in page ‘ $p$ ’ =  $\log_n(\alpha/\text{OutDegree}(p))$ , where,

$1/\text{OutDegree}(p)$  = probability of a random surfer in page ‘ $p$ ’ clicking on one of the links in page ‘ $p$ ’ and  $\alpha$  is a damping factor. An average click is one click among ‘ $n$ ’ links. The value of ‘ $n$ ’ is usually set to 7 since an average page has roughly 7 hyperlinks to other pages.

Summing the length of the links on a path is equivalent to multiplying the probabilities of traversing the links on a path. This leads to the definition of distance between two pages:

*“The distance between two pages  $p$  and  $q$  is defined as the sum of the lengths of the links on the shortest path from  $p$  to  $q$ ” [34].*

Average clicks measure can be used to filter Web sites in identifying communities within a certain distance. It can also be used in adaptive Web sites where the distance of a document from the root document is generally taken into account for minimizing the cost of a user to visit that document.

### **3.2.2 Information Scent**

*Information scent* is a concept that uses the snippets and information presented around the links in a page as a “scent” to evaluate the quality of content of the page it points to and the cost to access such a page [40]. The key idea is a user at a given page “foraging” for information would follow a link with a stronger “scent”. The “scent” of the pages will decrease along a path and is determined by network flow algorithm called *spreading activation*. The snippets, graphics, and other information around a link are referred as “proximal cues”. The user’s desired information is expressed as a weighted keyword vector. The similarity between the proximal cues and the user’s information need is computed as “Proximal Scent”. With the proximal cues from all the links and the user’s information need vector a “Proximal Scent Matrix” is generated. Each element in the matrix reflects the extent of similarity between the link’s proximal cues and the user’s information need. If enough information is not available around the link, a “Distal Scent” is computed with the information about the link described by the contents of the pages it points to. The “Proximal Scent” and the “Distal Scent” would then combine to give the “Scent” Matrix. The probability that an user would follow a link is decided by the “scent” or the value of the element in the “Scent” matrix. Chi et al. in [40] proposed two new algorithms called *Web User Flow by Information Scent (WUFIS)* and *Inferring User Need by Information Scent (IUNIS)* using the theory of *information scent* based on Information foraging concepts [40]. *WUFIS* tends to predict *user actions* based on *user needs* and *IUNIS* infers *user needs* based on *user actions*.

### 3.2.3 Bibliometric Metrics

The bibliometric community has defined a number of concepts around articles referencing other articles e.g. co-citation and co-reference. These concepts have been extended to the Web documents and have been briefly discussed in [2]. The measure of co-citation is basically the strength of the relationship between the two co-cited articles. The strength is measured by the count of the identical pages that cite the two pages together [35]. Similarly, measures for co-references, mutual references and indirect references can be developed. It is also possible for metrics for a single page to measure such concepts.

### 3.3 The Whole Web Graph

Some interesting Web graph properties are the diameter of the Web, the connectivity and the degree (indegree and outdegree) distributions. Albert et al [6] have defined and examined the diameter of the Web. Broder et al in [7,16] discovered that average diameter for a strongly connected component (see section 1.2.3) was at least 28. They also confirmed that the distributions of indegree and outdegree of pages follow a “power law” distribution, which suggests that the probability that a value  $d$  exists is proportional to  $1/d^p$ , where  $d$  belongs to a set of positive integers and  $p$ , the exponent, is some small positive number. The exponent for indegree was found to be approximately 2.1 and that for outdegree to be approximately 2.72. The average connected distance based on in-links was 16.12, and based on out-links was 16.18. For links without taking direction into account (undirected edges), the average connected distance was 6.83.

### 3.4 Other Related Measures

In [24], Cohn and Chang develop a “factored probabilistic model” of document citations called PHITS as a probabilistic analogue of the HITS approach. They assume a set of factors  $z$ , is given and they use it to determine the expectation that a particular document-citation ( $d$  and  $c$ ) pair is “explained” by  $z$ . The analog to “authority” is the probability  $P(c/z)$ . In a related paper [36] Cohn et al discuss methods to merge the PLSA (Probabilistic Latent Semantic Analysis) and the PHITS into a joint probabilistic model, explaining terms and citations by a common set of underlying factors. Borodin et al have made a comparative study of some of the hypertext link analysis algorithms in [26].



Amento et al. [37] have compare different link and content-based algorithms with rankings given by human experts on certain topics. They observed that simple in-degree performed as well as an authority or PageRank algorithm in the domain of their experiments, and according to the evaluation by the set of human experts. Ding et al. [38] analyze the HITS algorithm from the perspective of matrix algebra and a probabilistic approach. They show that in an average case there is a high correlation between the in-degree and the authority scores of HITS algorithm and the out-degree and the hub scores of the HITS algorithm. Zhu et al. [39] proposed a new metric called ‘PageRate’ to rate Web pages using usage data. No experimental results were provided and the metric becomes unstable in the boundary case when a page is pointed to by another page that has no incoming links.

## **4. Algorithms**

In this section, we survey the main algorithms used to compute the metrics described in section 3. These algorithms fall into two distinct categories. First consists of a set of basic algorithms, e.g. HITS, Maximal Flow Method that are basic approaches to analyzing Web graphs. Second are the approaches that use the basic algorithms as building blocks to compute more sophisticated metrics, e.g. the “Exact-Flow-Community” of Flake et al [15]. We do not discuss all versions and modifications of some of the base methods. We present here only the basic algorithms that have been most influential in hyperlink analysis and categorize them into those that apply to a single page or to multiple pages.

### **4.1 Algorithms for a single page**

#### **4.1.1 HITS (Hypertext Induced Topic Search) Algorithm**

The basis for the HITS algorithm is the concept of *hubs* and *authorities*. The main goal of the algorithm is to find the *hub* and *authority* scores of the Web pages related to a topic, which is used to identify the pages most relevant to the topic.

Let  $A$  be an adjacency matrix such that if there exists at least one hyperlink from page  $i$  to page  $j$ , then  $A_{i,j} = 1$ , else  $A_{i,j} = 0$ . Kleinberg's algorithm, popularly known as the HITS algorithm, is then run as follows:

### HITS ALGORITHM

Let  $\mathbf{a}$  is the vector of authority scores and  $\mathbf{h}$  be the vector of hub scores

$\mathbf{a}=[1,1,\dots,1]$ ,  $\mathbf{h} = [1,1,\dots,1]$  ;

**do**

$\mathbf{a}=\mathbf{A}^T\mathbf{h}$ ;

$\mathbf{h}=\mathbf{A}\mathbf{a}$ ;

Normalize  $\mathbf{a}$  and  $\mathbf{h}$ ;

**while  $\mathbf{a}$  and  $\mathbf{h}$  do not converge(reach a convergence threshold)**

$\mathbf{a}^* = \mathbf{a}$ ;

$\mathbf{h}^* = \mathbf{h}$ ;

**return  $\mathbf{a}^*$ ,  $\mathbf{h}^*$**

The vectors  $\mathbf{a}^*$  and  $\mathbf{h}^*$  correspond to the principal eigen vectors of  $A^T A$  and  $A A^T$ .

According to [17,33], the stability of the HITS algorithm to small perturbations is determined by the *eigengap* of  $S$ , which is defined as the difference between the largest and second largest eigenvalues. The HITS algorithm is less stable than Google's PageRank, and [17,33] proposes two new algorithms that are modification of the HITS algorithm, and have better stability.

The first algorithm, called *Randomized HITS*, introduces a bias factor based on time-step (odd or even) to determine authority and hub scores. It can be viewed as a random surfer tossing a coin with a bias,  $\epsilon$ . This bias is the probability that at any given time the surfer will jump to a new page chosen uniformly at random. With a probability  $1-\epsilon$ , the surfer will follow an out-link if it is an odd time-step and will traverse an in-link if it is an even time-step. The authority weight of the page is the chance that a surfer visits that page at an odd time-step  $t$ . The second algorithm is called *Subspace HITS*, that authority and hub scores are determined by the *subspace* spanned by the eigenvectors instead of individual eigenvectors. The bias factor or the subspace generated by the eigenvectors ensures more stability to perturbations than the original HITS algorithm.

### 4.1.2 PageRank Algorithm

The PageRank Algorithm is based on theory of random walk on a Markov Model. The algorithm involves computing the PageRank measure, mentioned in section 4, iteratively for the given set of pages. This can also be computed using matrix computations similar to HITS algorithm. The difference lies in the entries in the matrix. The matrix  $A$  used in a PageRank algorithm consists of transition probabilities. An  $(i,j)$  element in the matrix represents the probability that the link from page  $i$  to page  $j$  will be chosen. So for the initial values, the element  $(i,j) = 0$  if there is no link from page  $i$  to page  $j$ , else is  $1/OutDegree(i)$ , where  $OutDegree(i)$  is the outdegree of page  $i$  as defined in section 2.

#### THE PAGERANK ALGORITHM

Set  $\mathbf{PR} \leftarrow [r_1, r_2, \dots, r_N]$ , where  $r_i$  is some initial rank of page  $i$ , and  $N$  the number of Web pages in the graph;

$d \leftarrow 0.15$ ;  $\mathbf{D} \leftarrow [1/N, \dots, 1/N]^T$ ;

$\mathbf{A}$  is the adjacency matrix as described above;

**do**

$\mathbf{PR}_{i+1} \leftarrow \mathbf{A}^T * \mathbf{PR}_i$ ;

$\mathbf{PR}_{i+1} \leftarrow (1-d) * \mathbf{PR}_{i+1} + d * \mathbf{D}$ ;

$\delta \leftarrow \|\mathbf{PR}_{i+1} - \mathbf{PR}_i\|_1$

**while**  $\delta < \epsilon$ , where  $\epsilon$  is a small number indicating the convergence threshold

**return**  $\mathbf{PR}$ .

The vector  $\mathbf{PR}$  represents the global ranking of all the  $N$  Web pages in the Web graph.

## 4.2 Algorithms for Multiple Pages

Proposed algorithms that compute metrics for a set of pages that fall into two categories. First are the extensions of techniques for computing single page metrics, while the second are designed for multiple page metrics only. Those in the first category are the algorithms that are apply for a single page and extended to a group of pages. Thus, the basic algorithm remains the same. For example, the HITS algorithms is used to determine the *hub* and *authority* scores for a set of pages and can be used to identify communities [66,67] or for topic distillation [28]. Similarly, PageRank algorithm has been used for

identifying group of pages related to a topic[62]. The second are specifically designed for multiple pages, which will be the focus of this section. Other methods and procedures have been proposed that are both link and non- link based. For example, numerous search algorithms proposed for Web crawling include the Best First Search [43], Genetic Algorithms [44], FishSearch [45]. Techniques based on statistical methods, e.g. the Belief Propagation Method [46], EM algorithm [47] and the relaxation labeling techniques [48], have been used in classification of Web documents in [23] and [49]. These methods are additional tools used in Hyperlink analysis. Our study concentrates on the analysis of the information that explicitly takes advantage of the inter-document Web structure. And hence, we now discuss a method that has been designed for multiple pages.

#### 4.2.1 Maximal Flow Algorithm

Ford and Fulkerson [19] first suggested that equivalence of the maximal flow problem and the “minimum cut” problem in a graph. Some of the different implementations of the method proposed by them that is discussed in [42]. Here we just present the basic method by them. There are three important ideas that are come across in this method and other related flow algorithms:

- *Residual Network*: This is a network formed by a set of vertices of the original graph and a set of ‘residual edges’ that have some positive ‘residual’ capacity to allow for additional flow.
- *Augmenting path*: It can be described as simple path from a source  $s$  to a sink  $t$  in a residual network. It is a path from the source to the sink through which we can push some additional flow and then augment the flow along this path.
- *Cut*: A cut corresponds to partitioning the set of vertices,  $V$  of the graph into two sets of vertices, one containing the source  $s$  and the other containing the sink,  $t$ .

The Ford-Fulkerson-Method as described in [42] is s follows:

##### FORD-FULKERSON-METHOD

```

1      initialize flow  $f$  to 0
2      while there exists an augmenting path  $p$ 
3      do augment flow  $f$  along  $p$ 
4      return  $f$ 
```

On termination this process will return the maximal flow that will also correspond to the minimum cut. Flake et al [16], [22] have used the maximal flow approach to identify Web communities. The minimum cut will separate the Web graph into two sets. One set containing the *source* with the Web pages linked more densely among themselves, than to the other set containing the *sink*. The set of web pages that can be reached from the *source* represents the “Web Community”.

## 5. Analysis Scope

Hyperlink Analysis can be applied to some part of the Web graph or to the whole, depending on the needs of the application. We divide the scope of Hyperlink Analysis into three categories: *Single Page Analysis*, *Multiple Page Analysis* and *Whole Web Analysis*.

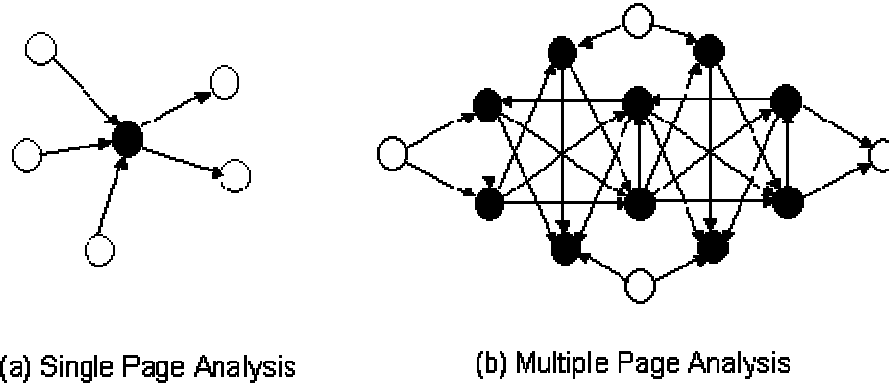


Figure 7: Kinds of Scope of Hyperlink Analysis

**Single Page Analysis:** The goal of such analysis is to understand the fundamental properties of individual pages and to derive quantitative metrics about them. For example, measures that define the properties of a page, such as PageRank, hub and authority scores, Web page reputations, etc., come under this category. Since links connect pairs of pages, other pages may be accessed to compute metrics about the page being focused on. However, the key is that the analysis itself is focused on individual pages.

**Multiple Pages Analysis:** The next type of analysis determines the properties of groups of pages. For example, the concepts of co-reference, co-citation, indirect reference, and mutual reference involve more than a single Web page. Similarly, the notion of “Web Community”, describing a group of related Web pages, also falls under this category. For applications whose analysis scope is multiple pages, it is also important to distinguish between single-site analysis and multiple-site analysis. While single-site analysis can be used to understand the structure of a given Web site, multiple-site analysis coalesces information from multiple Web sites, and thus presents more challenges such as data fusion, resource consumption, security, privacy, etc. Some of these issues may also exist for single-site analysis, particularly when the site being analyzed belongs to a different organization. For example, Liu et al. have used hyperlink analysis to discover unexpected information from a competitor’s Web site [51].

**The Whole Web Analysis:** At the highest level of analysis scope, the focus is to understand the properties of the entire Web. Deriving concepts and measures for characterizing such properties is an interesting research discipline not only because it helps to know how the Web has evolved over the past decade, but also because it may provide some hints as to what are the future requirements of Web applications, and how resource discovery systems can take advantage of the current structure of the Web. For example, Sherman [52] suggested that the bow-tie model proposed by Broder et al. [7]

“reveals a subtler structure that may lead to more efficient search engine crawling techniques and a greater understanding of the sociology of content creation, and that may help predict the emergence of new phenomena on the Web such as Web rings and spam clusters.”

Faloutsos et al [42] present a novel approach to study the structure of the Web based on power-laws relating the various properties of the Web graph such as their out-degree, frequency and rank of a node, and eigen values of the graph. They show that although the Web consists of autonomous entities, their topology is governed by simple power laws; and claim that such laws would continue to hold in the absence of revolutionary changes to the Internet. Knowing the existence of such power laws for describing the topological properties of the Web can help predict and extrapolate their future properties, and helps to design and analyze the performance of Web protocols.

## 6. Applications of Hyperlink Analysis

Hyperlink Analysis has been used in a wide variety of applications. These include determining the quality of Web pages related to a topic, classification of Web pages according to topics and other features, Web crawling, finding Web communities, building adaptive Web sites, personalization and page recommendations. Some of the main ones are now discussed in detail.

### 6.1 Topic Distillation

*Topic Distillation* is the identification of a set of documents or parts of document that are most relevant to a given topic. In [49] it has been defined as

*“the process of finding authoritative Web pages and comprehensive “hubs” which reciprocally endorse each other and are relevant to a given query.”*

Kleinberg ‘s HITS approach [29] was one of early link based approach that addressed the issue of identifying Web pages related to a specific topic. Bharath and Henzinger in [28] and Chakrabarti et al [27,30]) used hyperlink analysis to automatically identify the set of documents relevant to a given topic. Katz and Li in [53] use a three step approach – (i) *Document Keyword Extraction*, (ii) *Keyword propagation across pages connected by links*, and (iii) *keyword propagation through category tree structure* – to automatically distill topics from the set of documents belonging to a category or to extract documents related to certain topics. The FOCUS project [54, 55, 56] concentrates on building portals pertaining to a topic automatically. A “*fine-grained model*” based on the Document Object Model (DOM) of a page and the hyperlink structure of *hubs* and *authorities* related to a topic is described in [57,58]. This approach reduces *topic drift* and helps in identifying parts of a Web page relevant to a query. [59] describes a method of “spectral filtering” for resource discovery and [60, 61] has a compilation of interesting work concerning resource discovery and topic distillation.

In recent work on identifying topics, Rafiei and Mendelzon [13] define a new measure called “reputation” of a page and compute the set of topics for which a page will be rated

high. Haveliwala [62] proposed a “Topic-Sensitive PageRank”, which pre-computes a set of PageRank vectors corresponding to different topics.

## 6.2 Web page Categorization

*Web page categorization* determines the category or class a Web page belongs to, from a pre-determined set of categories or classes. *Topic Distillation* is similar but in Web page categorization, the categories can be based on topics or other functionalities, e.g. home pages, content pages, research papers, etc, whereas *Topic Distillation* is exclusively concerned mainly with content-oriented topics. Pirolli et al [63] defined a set of 8 categories for nodes representing Web pages and identified 7 different features based on which a Web page could be categorized into these 8 categories. Chakrabarti et al [49] use a *relaxation labeling technique* to model a class-conditional probability distribution for assigning a category by looking at the neighboring documents that link to the given document or linked by the given the document. Attardi et al [64] proposed an automatic method of classifying Web pages based on the link and context. The idea is that if a page is pointed to by another page, the link would carry certain context weight since it induces someone to read the given page from the page that is referring to it. Getoor et al [23] treat documents and links as entities in an Entity- Relationship model and use a *Probabilistic Relational Model* to specify the probability distribution over the document-link database, and classify the documents using *Belief Propagation* [46] methods. In recent work [65], Chakrabarti et al describe how topic taxonomies and automatic classifiers can be used to estimate the distribution of broad topics in the whole Web.

## 6.3 Identification of Web Communities

A “Web Community” can be defined as a group of pages that address similar topics or reflect the common interests of the creators of these pages. The similarity can be based on content as well as the inherent link structure [65]. However, this description of a web community is by no means a definition since various researchers have defined in different ways.

In [65], Gibson et al first developed the concept of “Web Communities” based on link structure. They defined a community as “a core of central ‘authoritative’ pages linked



together by ‘hub’ pages.” They used the HITS algorithm to identify the “communities” as defined by them. In [67], Ravi Kumar et al. described a process called *trawling*, which is described as “a systematic enumeration of emerging Web communities from a Web crawl”. This work concentrates on identifying community cores in the Web graph during a crawl. These cores can be used to then identify the relevant communities using the HITS approach. Identifying such cores indicate the possible emergence of communities that have not been recognized by human experts. In [16,22], Flake et al, defined Web communities as, “a collection of Web pages such that each member page has more hyperlinks (in either direction) within the community than outside of the community”. The base method to identify communities was the “maximal flow- minimum cut” method [19]. In [68], Adamic and Adar, introduced the concept of “Friends and Neighbors”. Instead of concentrating on Web pages for a topic, they focused their method on individual Web pages. Their goal was to identify a group of individuals with similar interests, who in the cyber-world would form a “community”. Two people are termed “friends” if the *similarity* between their Web pages is high. The *similarity* is measured using the features: *text*, *out-links*, *In-Links* and *Mailing Lists*.

Matsuura, Ohsawa and Ishizuka in [25] describe a method to identify topics that could interest people from multiple communities and could possibly grow into a topic of interest leading to form more communities. Thus their definition of a “community” is a “group of people sharing some value”. They define two kinds of pages:

**Archive Pages:** Pages of highest rank according to Google in a community.

**Agora Pages:** Pages linked from multiple archive-pages but are not in any community themselves are taken as novel topics attracting multiple communities, called *agora-topic* pages.

Thus these set of *Agora* pages reflect the emerging communities. In a more recent paper, Argyros et al [69] talk about extracting communities through patterns.

## 6.4 Web Crawling

The area of Web crawling and searching has become an interesting research field as the size of the Web is increasing. Given the size of the Web, it has become important to first search / crawl the Web pages relevant to the area of interest. Some of the early work in

determining the quality of a page to help determine the pages to crawl and other crawling methods can be found in [29, 31, 70, 71, 72, 73, 74, 75].

“*Focused Crawling*” is one of the methods proposed by Chakrabarti et al in [56, 57] for efficiently crawling pages that are associated with a topic. It is necessary to identify good *hub* pages that serve as a source of outgoing links for *authority* pages while crawling. Finally the *crawler* determines dynamically the links to be traversed and collects the necessary information [57]. The relevance of context in Web search and focused crawling is discussed in [76,77].

Menczer et al [78] propose three different methods to evaluate “*topic-driven Web Crawlers*”. The first approach is based on determining the quality of the pages crawled using a classifier. The second approach, they used an independent retrieval system; SMART for ranking the pages for the topic concerned. This is compared to the order in which the pages were crawled. A good crawler must crawl the pages in the descending order of the ranks, i.e. the high ranked pages must be crawled first. The third approach, takes the average cosine similarity of the *TF.IDF* vector of the topic and the *TF.IDF* vector of each page. If the crawler is seeking pages having “high –similarity” with the topic each time, then the crawl is said to be effective. In their experiments they compared *BestFirst* crawler, the *PageRank* crawler [71] and the *InfoSpiders* [79] crawler. Interestingly, they found that the naïve *BestFirst* performed better under their methods of evaluation.

Aggarwal et al. [80] proposed an “intelligent crawling” method. The approach takes into account the linkage structure of the Web and the following features: the content of the page, URL tokens like certain keywords that indicate the importance of a “candidate URL” with respect to a topic, the nature of inlinking of the Web pages of a given candidate URL, and the number of “siblings” (Web pages that have been co-cited) that have already been crawled. These features are used to determine the “Priority Value” according to which the pages will be crawled. Chakrabarti et al [81] suggest a “critic-apprentice” model to improve on their earlier focused crawling technique. The new

model consists of the classifier of the baseline-focused crawler developed earlier in [56, 57] and an “*apprentice*” that essentially is a learner and helps the baseline classifier to improve the crawling as it learns. Pant et al. [82] investigate the idea of exploring the links that are sub-optimal to lead to obtain more relevant pages in a topic specific crawling.

Web Crawling can also be used for business related applications e.g. Shah et al [83] use to crawl the e-bay Website and find out the history of auctions to study bidding strategies. Liu et al [51] use crawling to find “unexpected” information from competitors’ Websites and present it to the user.

## **6.5 Web Usage Based Applications**

Usage statistics can be combined with the link structure to produce interesting results.

Usage statistics has been applied to hyperlink structure for better link prediction for *adaptive Web sites*. Some of the early work on *adaptive Web sites* was done by Pekrowitz and Etzioni in [50, 84] and Cooley et al in [85]. [84] discusses predicting user-browsing behavior based on past surfing paths using Markov models. In [86] Sarukkai proposed the use of *link prediction* and *path analysis* for better user navigation. He proposes a Markov chain model to predict the user access pattern based on the user access logs previously collected. Zhu et al. [18] introduce the *maximal forward reference* approach to that of Sarukkai [86] to eliminate the effect of backward references by the user. Mobasher et al [87] have used usage statistics on the basic link structure for automatic personalization of the Web.

## **7. Methodology for Applying Hyperlink Analysis**

Based on the discussion so far, research on Hyperlink Analysis can be classified using the dimensions of *Knowledge Model*, *Metrics*, *Analysis Scope* and *Algorithms*. In our observation, these form the building blocks for Hyperlink Analysis. In this section, we first classify a number of applications reported in the literature based on these dimensions, and then propose a general methodology for applying hyperlink analysis to suit the purposes of an application.

## 7.1 Classification of Applications Using Hyperlink Analysis

Hyperlink analysis has been applied successfully to a number of applications. There are a lot of commonalities between them, even though it is not reported as such in the original sources. In Table 1, we summarize the approach taken by many of these projects, their application focus and the choices each approach has made on each of the four dimensions.. We should also note here that Hyperlink Analysis is often used as a technique to improve upon information mined from the content of web pages. Hence, in the table we have also included projects that use Web content as the primary source of data for the mining process. For example, the FOCUS [52] project uses text based *classifiers* and *distillers* that serve as portals to many relevant pages. The *distillers* therefore must be good hubs, and hence for crawling purposes it is essential to crawl good hubs first. These hubs themselves are determined using the HITS algorithm, a popular hyperlink analysis technique.

## 7.2 Hyperlink Analysis Methodology

The methodology for using hyperlink analysis for an application can be described as the following sequence of steps:

1. Analyze the needs of the application to determine the type of information it needs from hyperlink analysis. For example, the web search application requires that pages that are relevant to a user query be ranked in some order of importance. The information model here is a ranked list of URLs. In some cases a process model may be required in addition to the information model.
2. Next, determine the metric(s) that need to be calculated to quantify various aspects of the information model. For example, for Google the metric is PageRank, while in the HITS approach it is HubScore and AuthorityScore. As newer applications of hyperlink analysis are being discovered, new metrics will have to be developed to suit their needs.
3. Algorithms to compute the selected metrics need to be selected/designed next. The Google approach uses a (bounded scope) graph traversal algorithm to compute the PageRank metrics of pages relevant to a user query. The HITS

approach has developed a new algorithm called Hypertext Indexing and Topic Selection (HITS), which uses the algorithm for computing eigen values of large sparse matrices as its principal workhorse. Hyperlink analysis metrics and algorithms for computing them are intimately tied together, and each time a metric is designed, there will usually be a need to design an algorithm for it.

Project	Application Focus	Hyperlink Analysis				Web Content	Web Usage
		Knowledge Models	Metrics	Algorithms	Analysis Scope		
Google [1]	Web Search	Markov Model of random walk	PageRank	PageRank Algorithm	Single page	X	
HITS [29]	Web Search, Topic Distillation	Hubs and Authorities	Hub/Authority Scores	HITS Algorithm	Single page	X	
Bharat and Henzinger [28]	Topic distillation	Hubs and Authorities	Hub/Authority Scores	HITS	Single Page	X	
Clever [30]	Web Classification, Web Communities, Modeling Web Graph	Hubs and Authorities	Hub/Authority Scores	HITS	Single Page or Set of Pages	X	
FOCUS [54,55,56]	Web Crawling	Hubs and Authorities	Authority Scores	HITS	Set of Pages	X	
TOPIC [13]	Topic Distillation, Web page Classification	Markov Model of random walk + Hubs and Authorities	Web page reputation	PageRank Algorithm	Single Page, Set of pages	X	
SALSA [12]	Web Search	Hubs and Authorities and Markov chains	Modified Hub and Authority Scores	HITS	SinglePage	X	
Agora [25]	Web Communities	Agora Pages		PageRank and HITS	Set of Pages	X	
Friends and Neighbors [68]	Web Communities	InLinks, OutLinks, text	"similarity"		Set of two or more Home	X	
Flake et al [15, 22]	Web Communities, Web Crawling	Maximal-flow, minimal cut	Maximum density linkage	Maximal Flow Algorithm	Set of pages	X	
ONE [18]	Link Prediction	Markov Chains		Maximal Forward Method	Set of pages, Entire web		X
Matsuo et al[34]	Web Communities, Web Search	Markov Model	Average Clicks	Best First Search	Set of pages, entire web		
Getoor et al [23]	Web Classification	Probabilistic relational Model		Belief propagation	Set of pages	X	
PHITS[24,36]	Web Search, Web Classification	Probabilistic factored model	Maximal Likelihood		Set of pages	X	
Chakrabarti [57]	Topic Distillation	Hubs and Authorities including DOM of a page	Hub/Authority Scores	HITS	Set of pages	X	
Haveliwala [62]	Topic Distillation	Topic-Sensitive Page Rank	PageRank	HITS	Set of Pages	X	
Pirroli et al [63]	Web Classification	Inlinks, Outlinks			Single Page, Set of pages	X	

Table 1. Classification of Research Projects in Hyperlink Analysis.

4. Next, the analysis scope relevant to the application must be decided. The choices are single page level, groups of pages and links, or an entire graph. Similar analysis can be done with varying scopes for different applications.
5. Finally, it must be decided if hyperlink analysis is to be done just by itself, or in conjunction with web content and web usage analyses. If so, then the results must be integrated with those of the other analyses.

We believe that following this approach can help in better leveraging the growing body of techniques and experiences with hyperlink analysis.

## **8. Conclusions**

In this report, we have given a brief introduction to what *Hyperlink Analysis* means and its scope with regard to Web Mining. The variety of applications of this kind of analysis in the Web domain has led to rapid interests in this area. This has resulted in the development of a significant body of literature, reporting on emerging techniques for hyperlink analysis as well as experience in their usage.

An effort has been made to systematically catalog the existing research literature and bring out the similarities and complementarities of the different approaches. Four key dimensions, *Knowledge Models*, *Metrics*, *Algorithms* and *Analysis Scope* have been identified that help in classifying the research in the field of Hyperlink Analysis. For each dimension we list out the items that fall into the category from the existing research. We also give a description about the scope of Hyperlink Analysis and at what levels it can be applied to. The various applications of Hyperlink Analysis and the research that has gone into them is discussed in detail to give an idea of these different dimensions have been applied.

Finally, we present a methodology for applying Hyperlink Analysis for an application and the existing literature has been classified accordingly to see how the methodology can be implemented in the future. The state-of-art in the field is presented and key dimensions for methodology of Hyperlink Analysis has been identified. We hope this study serves as a good base for structuring future research in this area.

## 9. Acknowledgements

We thank the members of the Scout group at the AHPCRC at the University of Minnesota, for their valuable feedback and ideas presented in this paper. The Scout group consists of B. Uygur Oztekin, Levent Ertoz, Eric Eilertson, Aleksander Lazarevic, Kashif Riaz and Saurabh Singhal.

This work was partially supported by NSF grant number ACI 9982274, and by Army High Performance Computing Research Center contract number DAAD19-01-2-0014. The content of this work does not necessarily reflect the position or policy of the government and no official endorsement should be inferred. Access to computing facilities was provided by AHPCRC and the Minnesota Supercomputing Institute.

## References

- [1] <http://www.google.com>
- [2] Kemal Efe, Vijay Raghavan, C. Henry Chu, Adrienne L. Broadwater, Levent Bolelli, Seyda Ertekin (2000), *The Shape of the Web and Its Implications for Searching the Web*, International Conference on Advances in Infrastructure for Electronic Business, Science, and Education on the Internet- Proceedings at <http://www.ssgrr.it/en/ssgrr2000/proceedings.htm>, Rome. Italy, Jul.-Aug. 2000.
- [3] J. Srivastava, R. Cooley, M. Deshpande, and P. -N. Tan. *Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data (2000)*, SIGKDD Explorations, Vol. 1, Issue 2, 2000.
- [4] R. Kosala and H. Blockeel, "Web Mining Research: A Survey," SIG KDD Explorations, vol. 2, pp. 1--15, July 2000
- [5] Monika Henzinger, *Link Analysis in Web Information Retrieval*, ICDE Bulletin Sept 2000, Vol 23. No.3
- [6] R. Albert, H.Jeong, and A.-L. Barabasi. *Diameter of the World Wide Web*, Nature 401: 130-131, Sep 1999.
- [7] A. Broder et al, *Graph Structure in the Web*. In the Proc. 9<sup>th</sup> WWW Conference 2000.
- [8] O. Etzioni *The world wide web: Quagmire or goldmine*. Communications of the ACM, 39(11):65-68, 1996
- [9] Robert Cooley, Bamshad Mobasher, and Jaideep Srivastava, *Web Mining: Information and Pattern Discovery on the World Wide*, In Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97), Nov 1997.
- [10] S. Chakrabarti. *Data Mining for hypertext: A tutorial survey*. ACM SIGKDD Explorations, 1(2): 1-11, 2000.
- [11] L. Page, S. Brin, R. Motwani and T. Winograd (1998) *The PageRank Citation Ranking: Bringing Order to the Web*. Stanford Digital Library Technologies, Working Paper 1999-0120 available at <http://www-db.stanford.edu/~backrub/pageranksub.ps>, January 1998.
- [12] R. Lempel and S. Moran. *The stochastic approach for link-structure analysis (SALSA) and the TKC effect*. In the 9<sup>th</sup> International World Wide Web Conference, May 2000.
- [13] D. Rafiei, A.O. Mendelzon (2000), *What is this Page Known for? Computing Web Page Reputations*, In Proceedings of Ninth International WWW Conference, Amsterdam, May 2000.
- [14] D. Gibson, J. Klienber, and P. Raghavan. *Inferring web communities from link topology*. In Proc. 9<sup>th</sup> ACM Conference on Hypertext and Hypermedia, 1998

- [15] Gary William Flake, Steve Lawrence, C. Lee Giles . *Efficient Identification of Web Communities*. Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. August 2000, pp. 150-160
- [16] J. Klienberget al. *The web as a graph: measurement models and methods*. In Proc ICCD 1999.
- [17] A. Y. Ng, A. X. Zheng, and M. I. Jordan (2001), *Stable algorithms for link analysis*. Proc. 24th International Conference on Research and Development in Information Retrieval (SIGIR), 2001.
- [18] Jianhan Zhu, Jun Hong, and John G. Hughes, *Using Markov Chains for Link Prediction in Adaptive Web Sites*. In Proc. of ACM SIGWEB Hypertext 2002.
- [19] L.R. Ford Jr. and D.R. Fulkerson. *Maximal Flow through a network*. Canadian J. Math.,8:399-404, 1956
- [20] J. Edmonds and R.M. Karp. *Theoretical improvements in the algorithmic efficiency of network flow problems*. JACM, 19:248-264, 1972.
- [21] A.V. Goldberg and R.E.Tarjan, *A new approach to the maximal flow problem*, In Proc. 18<sup>th</sup> Ann. ACM Symposium on Theory of Computing, 1986
- [22] Gary William Flake, Steve Lawrence, C. Lee Giles, Frans M. Coetzee. *Self-Organization and Identification of Web Communities*. IEEE Computer, 35(3), 66–71, 2002.
- [23] L.Getoor, E.Segal, B.Tasker, D.Koller. *Probabilistic Models of Text and Link Structure for Hypertext Classification*. IJCAI Workshop on "Text Learning: Beyond Supervision", Seattle, WA, August 2001.
- [24] David Cohn and Huan Chang (2000). *Probabilistically Identifying Authoritative Documents*, Proceedings of the Seventeenth International Conference on Machine Learning, Stanford, CA
- [25] N. Matsumura, Y. Ohsawa, M. Ishizuka, *Discovering seeds of New Interest Spread from Premature Pages Cited by Multiple Communities*, In Proc. Of First Asia-Pacific Conference, Web Intelligence, Japan, October 2001.
- [26] A. Borodin, G.O. Roberts, J.S. Rosenthal, P. Tsaparas (2000), *Finding Authorities and Hubs From Link Structures on the World Wide Web*, WWW10 Proceedings , Hongkong, May 2001
- [27] Soumen Chakrabarti, B. Dom, D. Gibson, J. Kleinberg, P. Raghavan, and S. Rajagopalan. *Automatic Resource Compilation by Analyzing Hyperlink Structure and Associated Text*. Proceedings of the 7th World-Wide Web conference, 1998. Copyright owned by Elsevier Sciences, Amsterdam.
- [28] K.Bharat and M.R. Henzinger, *Improved Algorithms for topic distillation in hyperlinked environments*. In Proceedings of the 21<sup>st</sup> International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 104-111, 1998
- [29] J.M. Kleinberg, *Authoritative Sources in Hyperlinked Environment*, In Proc. Of the Ninth Annual ACM-SIAM Symposium on Discrete Algorithms, pages 668-667, 1998
- [30] The CLEVER project, IBM, <http://www.almaden.ibm.com/cs/k53/clever.html>
- [31] S. Brin and L.Page, *The anatomy of a large-scale hypertextual Web Search engine*. In 7<sup>th</sup> International World Wide Web Conference, Brisbane, Australia, 1998.
- [32] T. Haveliwala, *Efficient Computation of PageRank* In Technical Report, Stanford University, CA, 1999.
- [33] A. Y. Ng, A. X. Zheng, and M. I. Jordan (2001), *Link Analysis, Eigenvectors and Stability*, IJCAI-01
- [34] Y. Matsuo, Y.Ohsawa and M. Ishizuka (2001), *Average-clicks: A new measure of distance on the WWW*. In Proc. Of First Asia-Pacific Conference, Web Intelligence, Japan, October 2001.
- [35] Henry Small, *Co-Citation in the Scientific Literature: A New Measure of the relationship between two documents*, Essays of an Information Scientist, Vol:2, p.28, 1974-1976.
- [36] David Cohn and Thomas Hofmann. *The Missing Link - A Probabilistic Model of Document Content and Hypertext Connectivity*, In Advances in Neural Information Processing Systems, Vol. 13. The MIT Press. 2001
- [37] B. Amento<sup>1</sup>, L. Terveen, and Will Hill (2000), *Does "Authority" Mean Quality? Predicting Expert Quality Ratings of Web Documents* , ACM 2000.
- [38] C. Ding, H. Zha, X. He, P. Husbands and H. Simon, *Analysis of Hubs and Authorities on the Web*, Workshop on Web Analytics, Second SIAM conference on Data Mining, April 2002.
- [39] J. Zhu, J. Hong, J.G. Hughes, *PageRate: Counting the Web Users' Votes*, Proceedings of the twelfth ACM conference on Hypertext and Hypermedia 2001
- [40] Ed H. Chi, Peter Pirolli, Kim Chen, James Pitkow. *Using Information Scent to Model User*



*Information Needs and Actions on the Web*. In Proc. of ACM CHI 2001 Conference on Human Factors in Computing Systems, pp. 490--497. ACM Press, April 2001. Seattle, WA.

[41] Pirolli, P. and Card, S.K. (1999) *Information Foraging*. Psychological Review 106(4) (pp 643-675)

[42] T.H. Cormen, C.E. Leiserson and R.C. Rivest, *Introduction to Algorithms*, McGraw-Hill, New York, NY, 1990.

[43] Richard E. Korf and David Maxwell Chickering, *Best First minmax search*, Artificial Intelligence, 84:299-337, 1996

[44] D.E.Goldberg. *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley, Reading, MA, 1989.

[45] P.M.E. De Bra and R. D. J. Post. *Searching for arbitrary information in the WWW: the fish search for Mosaic*. In second WWW conference 1994.

[46] J. Pearl. *Probabilistic Learning in Intelligent systems*. Morgan Kaufmann 1988.

[47] A. Dempster, N. Laird and D.Rubin. *Maximum Likelihood from incompleted data via the EM algorithm*. Journal of Royal Statistical Society, 39(Series B), 1-38.

[48] S. Chakrabarti, B. Dom and P.Indyk. *Enhanced hypertext categorization using hyperlinks*. In SIGMOD CONFERENCE ACM, 1998.

[49] S. Chakrabarti, *Integrating the Document Object Model with hyperlinks for enhanced topic distillation and information extraction*. In the 10th International World Wide Web Conference, Hong Kong, May 2001.

[50] M. Perkowitz and O. Etzioni, *Towards adaptive Web sites: conceptual framework and case study*. In the Proc. of the 8<sup>th</sup> International World Wide Web Conference, 1999.

[51] Bing Liu, Yiming Ma, Philip S. Yu. *Discovering Unexpected Information from Your Competitors' Web Sites*. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD-2001), San Francisco, CA; Aug 20-23, 2001

[52] Chris Sherman, *New Web Map Reveals Previously Unseen 'Bow Tie' Organizational Structure*, Newsbreaks and Conference Reports, May 22 2000. <http://www.infotoday.com/newsbreaks/breaks.htm>.

[53] V. Katz and W.-S.Li. *Topic Distillation on hierarchically categorized Web Documents*. In Proc. of the 1999 Workshop on Knowledge and Data Engineering Exchange, IEEE, 1999.

[54] The FOCUS project, <http://www.cs.berkeley.edu/~soumen/focus/>

[55] S. Chakrabarti, Martin van den Berg and Byron Dom, *Focused crawling: A new approach to topic-specific Web resource discovery*. In Proc. of 8<sup>th</sup> International World Wide Web Conference, Toronto, Canada, May 1999.

[56] S. Chakrabarti, Martin van den Berg and Byron Dom, *Distributed Hypertext Resource Discovery Through Examples*. In Proc. of 25<sup>th</sup> VLDB Conference, (Edinburgh, Scotland 1999), Morgan- Kaufman, pp, 375-386.

[57] S. Chakrabarti. *Integrating the Document Object Model with hyperlinks for enhanced distillation and information extraction*. In the 10<sup>th</sup> International World Wide Web Conference on the World Wide Web, Orlando, FL, Oct. 2001.

[58] Soumen Chakrabarti, Mukul Joshi, Vivek Tawde: *Enhanced Topic Distillation Using Text, Markup Tags, and Hyperlinks*, In the Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, September 9-13, 2001, New Orleans, Louisiana, USA. ACM 2001, ISBN 1-58113-331-6, pp 208-216

[59] S. Chakrabarti, B. Dom, S. R. Kumar, P. Raghavan, and A. Tomkins, *Spectral filtering for resource discovery*. In the Proc. of the SIGIR 98 Workshop on Hypertext Information Retrieval for the web, 1998.

[60] S. Chakrabarti. *Recent results in automatic Web Resource discovery*. ACM Computing surveys, Dec 1999.

[61] S. Chakrabarti, B.E. Dom, D. Gibson, R. Kumar, P. Raghavan, S. Rajagopalan, and A.Tomkins. *Experiments in Topic Distillation*. In Proc. of the ACM SIGIR Workshop on the Hypertext Information Retrieval on the Web, Australia, 1998.

[62] T. Haveliwala, *Topic-Sensitive PageRank*. In Proc. of the 11<sup>th</sup> International World Wide Web Conference, Honolulu, Hawaii. May 2002.

[63] P.Pirolli, J. Pitkow, R.Rao. *Silk from a sow's ear: Extracting usable structures from the web*. In the Proc. of ACM SIGCHI Conference on Human Factors in Computing. 1996.

[64] Giuseppe Attardi, Antonio Gulli, Fabrizio Sebastiani, *Automatic Web Page Categorization by Link*

and Context Analysis, Proceedings of THAI-99, 1st European Symposium on Telematics, Hypermedia and Artificial Intelligence. 1999.

[65] Soumen Chakrabarti, Mukul M. Joshi, Kunal Punera, David M. Pennock, *The Structure of Broad Topics on the Web*, In the 11<sup>th</sup> International World Wide Web Conference, Honolulu, Hawaii, May 2002.

[66] D. Gibson, J. Klienber, and P.Raghavan. *Inferring web communities from link topology*. In Proc. 9<sup>th</sup> ACM Conference on Hypertext and Hypermedia. 1998.

[67] Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, and Andrew Tomkins. *Trawling the web for emerging cyber-communities*. In Proc 8<sup>th</sup> Int. World Wide Web Conf., 1999

[68] L. A. Adamic and E. Adar, *Friends and Neighbors on the Web*, Xerox Palo Alto Research Center Palo Alto, CA 94304. 2000.

[69] T. Argyros, Ch. Ermopoulos, B. Paulaki, N. AlSaid, *Extracting cyber communities through patterns*, June 2002.

[70] I. Ben-Shaul, M. Herscovici, M. Jacovi, Y. Maarek, D. Pelleg, M. Shtalhaim, V. Soroka, and S. Ur. *Adding support for dynamic and focused search with fetuccino*. In Proceedings of 8<sup>th</sup> International World Wide Web Conference, 1999.

[71] J. Cho, H. Garcia-Molina, and L. Page. *Efficient crawling through url ordering*. In Proc. of the Seventh International World Wide Web Conference, 1998.

[72] H. Chen, Y. -M. Chung, M. Ramsey and C.C. Yang, *A smart itty bitsy spider for the web*, J. Am. Soc. Inf. Sci. 49(7) (1998) 604-618.

[73] P. Debra and R. Post, *Information Retrieval in World Wide Web: making client based searching feasible*, in : Proc. of the 1<sup>st</sup> International World Wide Web Conference, Geneva, Switzerland, 1994

[74] B. Pinkerton. *Finding what people want: Experiences with the webcrawler*. In the Proc. of the First International World Wide Web Conference, Geneva, Switzerland, 1994.

[75] M. Hersovici, M. Jacovi, Y.S. Maarek, D. Pelleg, M. Shtalheim and S. Ur. *The Shark Search algorithm – an application: tailored Web Site mapping*, In the Seventh World Wide Web Conference, April, 1998.

[76] M. Diligenti, F. Coetzee, S. Lawrence, C. Lee Giles, and Marco Gori. *Focused crawling using context graphs*. In the 26<sup>th</sup> International Conference on Very Large Databases, VLDB-2000, Cairo, Egypt 10-14 September 2000.

[77] S. Lawrence, *Context in Web Search*. IEEE Data Engineering Bulletin, Volume 23, Number 3, pp. 25-32, 2000

[78] F. Menczer, G. Pant, M. Ruiz, and P. Srinivasan. *Evaluating Topic Driven Web Crawlers*. In the Proc. 24<sup>th</sup> Annual Intl. ACM SIGIR CONF. On Research and Development in Information Retrieval.

[79] F. Menczer and R. Belew. *Adaptive retrieval agents: Internalizing local context and scaling upto the web*. Machine Learning, 39(2/3):203-242, 2000.

[80] Charu C. Aggarwal, Fatima Al-Garawi, Philip S. Yu, *Intelligent Crawling on the World Wide Web with Arbitrary Predicates*, In the Proceedings of 10<sup>th</sup> World Wide Web Conference, Hongkong, May 2001.

[81] S. Chakrabarti, K. Punera, M. Subramanyam, *Accelerated Focused Crawling through Online Relevance Feedback*. In the Proceedings of 11<sup>th</sup> World Wide Web Conference, Honolulu, Hawaii, May 2002.

[82] G. Pant, P. Srinivasan, F. Menczer: *Exploration versus Exploitation in Topic Driven Crawlers*. In the Workshop on Web Dynamics at the International World Wide Web Conference , Honolulu, Hawaii May 2002.

[83] Harshit S. Shah, Neeraj R. Joshi, and Peter R. Wurman, *Mining for Bidding Strategies on eBay*, Workshop of WebKdd 2002, Edmonton, Canada, July 2002.

[84] P. Pirolli, J. E. Pitkow, *Distribution of Surfer's Path Through the World Wide Web: Empirical Characterization*. *World Wide Web 1*:1-17, 1999

[85] Bamshad Mobahser, Robert Cooley, Jaideep Srivastava, *Creating Adaptive Web Sites Through Usage-Based Clustering of URLs (1999)*. Proceedings of the 1997 IEEE Knowledge and Data Engineering Exchange Workshop (KDEX-99), November 1999.

[86] R.R. Sarukkai, *Link Prediction and Path Analysis using Markov Chains*, In the Proc. of the 9<sup>th</sup> World Wide Web Conference , 1999

[87] Bamshad Mobasher, Robert Cooley, Jaideep Srivastava, *Automatic Personalization Based on Web Usage Mining*. Communications of the ACM, Vol. 43, No. 8, 2000.